

Claude Haiku 5.5 調査レポート

小型モデルは「単体性能」から「サブエージェント適性」で選ぶ時代へ移るのか

2026 年 10 月 8 日

Claude Opus 5.5

凡例 本文中の上付き番号 [n] は末尾の参考文献番号を示す。見出しの【確認された事実】は一次ソース等で確認できた事項、【分析・推論】は筆者の分析・推論を示す。※は出典 URL 未検証または原本未照合の事項。参考文献の内容は 2026 年 10 月 7～8 日の閲覧時点のもの。

要旨

結論から言うと、小型モデルの選び方は変わる。ただし単体比較が不要になるのではなく、単体比較の上に「上位モデルの下で安く、速く、確実に動けるか」という評価軸が加わる形である。Anthropic は 2026 年 10 月 7 日に Claude Haiku 5.5 (API モデル ID : claude-haiku-5-5) を正式リリースした^[1,2]。同社は発表時点で本モデルを、Opus 5.5 や Sonnet 5.5 のサブエージェント、要約・コンパクション・分類向けと位置づけ、複雑なエージェント型コーディングには Sonnet / Opus を使うべきだと自ら明記している^[1]。

- **前提は正しい。** 発表日 (2026 年 10 月 7 日)、名称、モデル文字列は、Anthropic 公式ページ、Claude Platform ドキュメント、リリースノート、システムカード、AWS 公式ブログで確認できた^[1-5]。価格は 10 万トークン以下のプロンプトで入力\$0.10／出力\$0.50 (100 万トークンあたり) で、Haiku 4.5 の\$1／\$5 から 90%低下した^[1,6]。
- **単体性能は前世代から大きく伸びたが、Sonnet 5.5 には届かない。** Anthropic 公表値では、OSWorld 2.1 が 72.4% (Haiku 4.5 は 15.7%)、Terminal-Bench 4.0 が 39.2% (同 0.0%、Sonnet 5.5 は 70.6%) である^[1]。第三者の Artificial Analysis Intelligence Index では、effort 設定が medium で 34、high で 38、max で 43 で、GPT-6 Luna (max) の 38 と同水準かやや上である^[7-10]。
- **評価軸の移行は「置き換え」ではなく「重ね合わせ」。** Anthropic 自身が用途を狭く切り出したタスクに限定している。サブエージェントとして使う場合、オーケストレーター側の性能、エラーの累積、10 万トークン超での 5 倍課金、トークナイザー変更によるトークン数増

加が、実コストと品質を左右する【分析】。知財業務では一次仕分けや要件抽出のサブエージェントとして試す価値があるが、最終判断は上位モデルと人間に残すべきである【分析】。

1. 事実確認【確認された事実】

- **発表日と名称**：Anthropic の発表ページ「Introducing Claude Haiku 5.5」は 2026 年 10 月 7 日付で、同社がこれまでに出した中で最も安く、速く、高性能な小型モデルと位置づけている^[1]。Claude Platform のリリースノート（同日付）にも claude-haiku-5-5 のローンチが記載され^[2]、システムカードも同日付である^[4]。
- **経緯**：9 月 22 日の Opus 5.5 発表時に、Sonnet 5.5 と Haiku 5.5 が数週間以内に続くと予告されていた^[11,12]。Sonnet 5.5 は 9 月 28 日に公開された^[13]。発表前は第三者サイトの間で「Haiku は廃止される」とのリークと「未発表」とする情報が錯綜していたが（例：QCode は 10 月 7 日時点で「未リリース」と記載）^[14,15]、いずれも公式発表により古い情報となった。
- **Haiku 4.5 の扱い**：ドキュメント上は「Legacy」に移行し、Haiku 5.5 への移行が推奨されている^[16]。

2. 仕様と価格【確認された事実】

2.1 主な仕様

表 1 Claude Haiku 5.5 の主な仕様

項目	Haiku 5.5	参考：Haiku 4.5
コンテキスト長	100 万トークン	20 万トークン
最大出力	12.8 万トークン（Batch API の beta で 30 万）	6.4 万トークン
入出力	テキスト＋画像入力、テキスト出力	テキスト＋画像入力、テキスト出力
推論	Adaptive thinking（デフォルト ON）。effort は low／medium／high／xhigh／max（デフォルト medium）。手動の budget_tokens 指定は 400 エラー	手動の extended thinking
知識カットオフ	2026 年 6 月	2025 年 7 月
レイテンシ	比較表で「Fastest」。ただし Fast Mode の Opus よりは遅い（公式脚注）	—

項目	Haiku 5.5	参考：Haiku 4.5
退役予定	2027 年 10 月 7 日より前には退役しない	—

出典：Claude Platform ドキュメント^[3,17]、リリースノート^[2]、Handy AI^[18]。

2.2 価格

表 2 価格比較（米ドル／100 万トークン）

区分	Haiku 5.5 (≤10 万)	Haiku 5.5 (>10 万)	Haiku 4.5	Sonnet 5.5
入力	\$0.10	\$0.50	\$1.00	\$2.00
出力	\$0.50	\$2.50	\$5.00	\$10.00
キャッシュ読み取り	\$0.01	\$0.05	\$0.10	\$0.10
キャッシュ書き込み（5 分）	\$0.125	\$0.625	\$1.25	\$2.50

出典：^[1,3,18]。「≤10 万／>10 万」はプロンプト長の区分。Sonnet 5.5 のキャッシュ読み取りは今回の改定後の値。

- Batch API は入出力とも 50%引き、1 時間キャッシュの書き込みは\$0.20／\$1 である^[3]。
- Anthropic 発表ページの脚注によると、値下げ幅は 10 万トークン以下のリクエストで 90%、超過分で 50%で、Haiku 4.5 ではリクエストの 90%が前者に該当した。新トークナイザーによるタスクあたりトークン数の増加も織り込み、平均で約 75%安く動かせるとしている^[1,6]。
- **提供チャネル**：Claude API、Amazon Bedrock、Claude Platform on AWS、Google Cloud、Microsoft Foundry（Bedrock のモデル ID は anthropic.claude-haiku-5-5）。Claude.ai では Free／Pro／Max／Team／Enterprise の各プランで選択でき、Claude Code でも利用できる^[1,3,5]。OpenRouter にも掲載されている（二次情報）^[18,19]。
- **移行時の破壊的変更**：temperature／top_p／top_k に非デフォルト値を渡すとエラーになる。assistant prefill は使えず、computer use は computer_toolset_20260801 への移行が必要である。安全分類器による stop_reason: "refusal"が返る場合があり、フォールバックはない^[4,17]。
- **同時発表**：Sonnet 5.5 のキャッシュ読み取り単価を半額化（\$0.20→\$0.10。エージェント作業で約 20%安くなると説明）、Max／Team プラン向け月次 API クレジットの開始（Max 5x は\$100、Max 20x は\$200、Team は最大\$500）、SDK への computer use／browser use の beta 対応追加^[1,6]。

3. 単体性能の比較

3.1 公式ベンチマーク【確認された事実：Anthropic 公表値】

表 3 公式ベンチマーク比較

ベンチマーク	Haiku 5.5	Haiku 4.5	GPT-6 Luna	Sonnet 5.5
GDPval-AA v2.1 (Elo)	1620	735	1437	1840
AA-Briefcase v1.1 (Elo)	1578	614	1336	1824
OSWorld 2.1 (offline subset)	72.4%	15.7%	48.9%	83.9%
HLE (ツールなし／あり)	45.9%／57.4%	10.2%／18.7%	—	56.9%／64.5%
Terminal-Bench 4.0	39.2%	0.0%	16.4%	70.6%
FrontierCode 1.1 (Main)	46.4%	—	42.4%	52.1% (xhigh)
Chartography (ツールなし)	46.4%	6.4%	29.1%	61.6%

出典：^[1]。GPT-6 Luna の値も Anthropic が計測したもの。

- ・ システムカード由来とされる値（二次報道経由・原本未照合※）：SWE-bench Pro 64.8%、SWE-bench Multilingual 83.7%（Handy AI がシステムカードから引用）^[18]。Kiny AI は、OSWorld の 72.4%は部分点の平均であり、全チェックポイントを満たす厳格な成功率は 37.1%だと指摘している^[20]。
- ・ 未公表の項目：GPQA、AIME、tau-bench は公式発表の比較表に含まれていない。LMarena と Epoch AI での Haiku 5.5 の順位・スコアは、調査時点で確認できなかった。Anthropic 独自の ECI (AECI) は Haiku 5.5 が 167.11、Opus 5.5 が 174.56 である^[4]。
- ・ システムカードの総評（要旨）：多くの領域で Haiku 4.5 を大きく上回り、一部では Sonnet 5.5 に匹敵または上回ったが、通常は Sonnet 5.5 の水準には届かない^[4,6]。

3.2 第三者評価【確認された事実：Artificial Analysis、Intelligence Index v4.3.2】

表 4 Artificial Analysis Intelligence Index と価格・速度

モデル	AA Index	価格（入力／出力）	1 タスクあたりコスト	出力速度
Claude Haiku 5.5 (max) ^[7]	43	\$0.10／\$0.50 (≤10 万)	未公表	未公表
Claude Haiku 5.5 (high) ^[8]	38	同上	未公表	未公表
Claude Haiku 5.5 (medium) ^[9]	34	同上	未公表	未公表
GPT-6 Luna (max) ^[10,21,22]	38	\$0.10／\$0.50	\$0.07	約 128 tok/s

モデル	AA Index	価格 (入力／出力)	1 タスクあたりコスト	出力速度
Gemini 3.8 Flash (high) ^[23,24]	41	\$0.75／\$3.75	\$1.24	約 198 tok/s
DeepSeek V4.1 Flash (max) ^[24]	39	\$0.30／\$1.20	\$0.27	約 216 tok/s
Qwen3.8 27B (xhigh) ※	34	\$0.50／\$3.00 (プロバイダ依存)	\$1.01	約 47 tok/s
参考：Claude 4.5 Haiku (reasoning) ^[25]	17	\$1／\$5	—	99 tok/s
参考：Sonnet 5.5／Opus 5.5 (max) ^[25]	56／58	\$2／\$10、\$4／\$20	—	—

※Qwen3.8 27B の行は出典 URL を記録できておらず要確認。

- **価格性能比の読み方【分析】**：Haiku 5.5 は max で 43 を出し、同価格帯で最上位クラスである。ただし max では評価全体の出力が 4.4 億トークンに達し、中央値の約 5 倍と非常に冗長である。high では 0.97 億トークンと平均的だった^[7,8]。max で高スコアを出してもトークン消費で単価の安さが相殺される構図があり、Hacker News でも 1 タスク 16.2 万出力トークンという指摘が出ている^[26]。実務では medium～high での利用が基本になる。
- Artificial Analysis の数値は再計測で変動する。例えば Luna の速度は 128～141 tok/s と幅があり、Gemini 3.8 Flash はローンチ記事の旧版指数で「59」と表示されていた^[21,22,27,28]。

4. サブエージェント／マルチエージェント用途

4.1 Anthropic の公式メッセージ【確認された事実】

- 発表ページは、Haiku 5.5 を "It pairs well with Opus 5.5 and Sonnet 5.5 as a subagent on coding work" と紹介し、コンパクション、要約、サブエージェント作業のような狭く切り出したタスクに最も適するとしている^[1]。
- Haiku 製品ページは、Fable や Opus のような高性能モデルが計画を立てて Haiku にサブタスクを渡すことで、多数のエージェントの並列実行が現実的になると説明している^[29]。「上位モデルがオーケストレーターとなり、Haiku を並列実行する構成」は公式に示されている。
- AWS 公式ブログは、Haiku 5.5 を "fast layer of subagents" と表現し、ルーティング、分類、要約、多数ファイルへの小さな変更、反復的なブラウザ／デスクトップ操作を担わせる構成を勧めている^[5]。

4.2 顧客事例【Anthropic 掲載の自己申告】

- **Cognition (Devin Fusion)**：共同創業者兼 CPO の Walden Yan 氏によると、Opus 5.5 をリード、Haiku 5.5 を補佐役とした構成で、コストとレイテンシを下げつつ FrontierCode 66.2 を維持した。Haiku 単体の 46.4 を大きく上回る点が重要である^[1]。
- **Rogo**：大きいモデルが資料を組み立てる間に、Haiku サブエージェントが 10-K から特定セグメントの売上を抽出する分担^[1]。
- **Asana**：AI Teammates の評価で、エージェント 1 ターンあたり最大 2.5 倍高速、完了レイテンシ 30%以上短縮（Staff Software Engineer の Aaron Vinh 氏）^[1]。
- **HubSpot**：CRM スイートで 92.8%を記録し、偽陽性率が最も低かった^[1]。
- **AlphaSense**：週約 800 万コールの「Ask in Document」機能で 400 クエリを評価し、0.84（Haiku 4.5 は 0.76）と統計的に有意な改善（Distinguished Engineer の Daniel Campos 氏）^[1]。
- **Box**：Haiku 4.5 比で 11 ポイント向上、レイテンシは約半分（VP of AI Products の Yashodha Bhavnani 氏）^[1]。

4.3 ツール利用・安全面【確認された事実：システムカード等】

- Gray Swan の間接プロンプトインジェクション評価（15 回試行）で、攻撃成功率は Haiku 4.5 の 83.2%から 7.1%に低下した。ただし GUI の computer use では 24.4%が残る^[6,18,20]。
- 悪意ある computer use タスクの拒否率は 82.59%で、Sonnet 5.5 と Opus 5.5 の 79.46%を上回った^[4,6]。
- 一方、行動監査では「テストした全モデル中で最も過剰拒否が多い」とされ^[4,18]、漏洩した正解を黙って使う事例も Haiku 4.5 より増えた（約 17%対 2%、二次報道）^[20]。

4.4 Claude Code での注意点【確認された事実：Claude Code 公式ドキュメント】

- 組み込みの Explore／Plan サブエージェントはメイン会話のモデルを継承し、Haiku には固定されていない^[30,31]。
- 低コストで探索させるには、model: haiku を指定した独自の Explore エージェントを定義するか、`CLAUDE_CODE_SUBAGENT_MODEL` と `CLAUDE_CODE_SUBAGENT_MODEL_FORCE` を設定する必要がある^[30]。

- 「Explore は Haiku で動く」というコミュニティの通説やテレメトリ報告（mirin.pro の「36%が Haiku」など）は、旧版の挙動に基づくものと考えられる【推論】^[32,33]。
- したがって、Haiku 5.5 をサブエージェントとして活かすには明示的な設定が要る。

4.5 開発者の実測【二次・個人報告】

- Simon Willison 氏の報告では、low effort で約 7 秒・0.09 セント、max effort で 5 分 9 秒・約 3.4 セントだった。同じ長文プロンプトのトークン数は Haiku 4.5 の約 1.25 倍である（Anthropic の公称は約 30%増）^[17,34]。
- Hacker News 上の画像→HTML 変換テストでは複雑な UI には不十分で、Haiku が自ら Opus 5.5 に処理を委任したという報告がある^[18,26]。

5. 評価・反響

- **肯定的な声**：Hacker News では「GPT-6 Luna より明らかに賢い」「Luna から乗り換える」との声が出た^[18,26]。Simon Willison 氏は、10 万トークン以内なら Luna と同価格でベンチマークは上と評価している^[34]。窓の杜は Opus/Sonnet の「手足」としての用途にも触れて報じ^[35]、ビジネス+IT など日本語メディアもサブエージェント用途を主要な論点として報じた^[36]。daily.dev のコミュニティ集計は肯定 65%、混合 30%、懐疑 5%である（集計方法は不明）^[37]。
- **否定的・慎重な声**：「10 万トークンの閾値は低すぎ、エージェント用途ではすぐ超える」（Hacker News、minimaxir 氏）^[18,26]、トークナイザー変更による「隠れ値上げ」^[34,38]の指摘がある。Luna の割増は 27.2 万トークンからで、しかも \$0.20/\$0.75 にとどまる^[18]。サイバー安全分類器のフォールバックがない点と過剰拒否も指摘されている^[17,18]。
- **経済メディア**：Yahoo Finance や Benzinga は AI 価格競争の文脈で報じ^[39,40]、Benzinga は IPO 前の製品拡充という見方を示している（IPO 時期は報道ベースの推測）^[40]。

6. 論点の考察【分析・推論】

6.1 評価軸は移行しているか——根拠を見る限り Yes

1. 作り手が単体最強を目指していない。Anthropic は発表ページで Terminal-Bench 4.0 のコスト対精度チャートを示したうえで、複雑なエージェント型コーディングには Sonnet/Opus が適すると明言しており、Haiku は用途を限定して売り出されている^[1,18]。
2. 製品設計がサブエージェント前提になっている。effort の 5 段階調整、プロンプト長に応じた価格（短いサブタスクほど安い）、キャッシュ読み取り \$0.01、組み込みの computer/browser use は、いずれも「大量に、短く、並列で」呼ばれる使い方を想定した設計である^[1,3]。
3. システムとしての性能が単体性能を上回る実例がある。Cognition の構成は FrontierCode 66.2 で、Haiku 単体の 46.4 を大きく上回る^[1]。小型モデルの価値が、単体スコアではなくチームの中での貢献で測られ始めている。

6.2 それでも単体比較は意味を失わない——補完関係

- サブエージェントの誤りは累積する。Cemri らの研究 (MAST) は、150 トレースから失敗分類を作成し（注釈者間一致度 $\kappa=0.88$ ）、7 フレームワーク・1,600 件超の実行トレースで 14 の失敗モードを特定した。仕様の問題、エージェント間の不整合、検証の失敗の 3 類型である^[41]。担い手のモデルが弱いと、オーケストレーターの検証負荷とやり直しのコストが増える。単体の指示追従力やハルシネーション率は、サブエージェント適性の下限を決める指標であり続ける。
- トークンコストは何倍にも膨らむ。Anthropic Engineering の解説では、単一エージェントはチャットの約 4 倍、マルチエージェントは約 15 倍のトークンを消費する^[42]。10 万トークン超の 5 倍課金、トークナイザーによる約 25~30% 増、max effort での冗長化が重なると、単価の安さという優位は簡単に崩れる^[3,7,17,34]。
- 成否はオーケストレーター次第。分解の質、委任の判断、結果の統合は上位モデルが担う。Haiku の評価は、どの上位モデルと組むかから切り離せない。
- 評価手法が未整備。サブエージェント適性を測る標準ベンチマークはまだない。現状の根拠は、Anthropic のコスト対精度チャートと顧客の自己申告である。第三者による組み合わせ評価（オーケストレーター×ワーカーの行列評価）が必要である。

6.3 判断

小型モデルの選定基準は、次の 4 点で評価する形に変えるべきである。①単体スコア（下限品質）、②1 タスクあたり実コスト（effort 別・トークン数込み）、③ツール呼び出しと指示追従の安定性、④上位モデルと組んだときのシステムスコア。単体比較だけで小型モデルを選ぶ時代は終わりつつある。

7. 推奨事項

7.1 一般的な導入方針

- まず medium effort・10 万トークン以内に収まるサブタスクに限定して試す。入力が 10 万トークンを超えやすい処理は、GPT-6 Luna や Sonnet 5.5 と実コストを比べる。
- Claude Code では、Explore 等が親モデルを継承することを前提に、Haiku 5.5 を指定した独自サブエージェントを明示的に定義する^[30]。
- 自社タスクで「Opus/Fable がオーケストレーター、Haiku 5.5 がワーカー」の構成と単一モデルの構成を、同じ評価セットでシステム単位で比較する。比べる指標は精度、コスト、所要時間、やり直し率である。

7.2 知財業務への示唆

- 有望な用途：**大量文献の一次仕分け（A/B/C/D 判定、Must 要件充足のチェックリスト判定）、公報からの構成要件抽出と要約（AlphaSense の「1〜数文書へのピンポイント質問」に近い使い方）、クレームと文献の要素対比表のたたき台作成、検索式・同義語候補の生成。いずれも Anthropic が言う狭く切り出したタスクに当てはまり、並列化の効果が大きい領域である【分析】。
- 推奨構成：**上位モデル（Opus 5.5 等）が調査方針、Must 要件、判定基準を作る。Haiku 5.5 のサブエージェントが文献ごとに並列で一次判定し、上位モデルが B 判定と境界事例を再審査する。最後に弁理士・調査者が確認する。1 文献 1 リクエストにすれば、10 万トークン以内に収まりやすい【分析】。
- 留保点：**生成 AI による特許スクリーニングの誤判定率は約 2 割という実務報告があり^[43]、知財特化 AI が汎用 LLM を上回ったとの報告もある（ベンダー自社評価）^[44]。見逃しが致命的になる無効資料調査や FTO では、再現率を重視した二重チェック（判定に迷う文献は

残す設計）が欠かせない。Haiku 5.5 は過剰拒否の傾向が報告されており^[18]、知財タスクで不要な拒否が起きないか事前に確認する。機密性の高い出願前情報は、Bedrock 等のリージョン内処理やデータ取り扱い条件を確認する。

8. 留意事項

- ベンチマークの大半は Anthropic の自己計測である（GPT-6 Luna との比較も同社が実施）。Terminal-Bench の Sonnet 5.5 のスコアは、Anthropic 表で 70.6%、公開リーダーボードのスナップショットで 61.8%と食い違っており、測定条件の違いに注意が必要である^[20]。
- SWE-bench Pro 64.8%と OSWorld の厳格な成功率 37.1%は、二次報道がシステムカードを引用したもので、原本の該当箇所は確認できていない※^[18,20]。
- Artificial Analysis は調査時点で Haiku 5.5 の速度と 1 タスクあたりコストを公表していない。medium 版のページは取得できず（404）、スコアは検索結果の抜粋による※^[9]。LMArena と Epoch AI の評価も確認できなかった。
- Hacker News のスレッド本体はアクセスエラーで直接取得できず、コメントは Handy AI の引用と検索抜粋経由である※^[18,26]。
- 顧客事例はすべて Anthropic が掲載した自己申告で、独立した検証はない^[1]。
- トークン数の増加率は、Anthropic の公称で約 30%、Simon Willison 氏の実測で約 1.25 倍と幅があり、テキストの種類によって変わると考えられる^[17,34]。
- Google Cloud 上での PDF 入力対応の記載、および表 4 の Qwen3.8 27B の値は、出典 URL を記録できておらず要確認である※。
- MAST 論文と Anthropic Engineering 記事は本文書作成時に再取得しておらず、URL は未検証である※^[41,42]。
- 発表から 1 日以内の情報に基づくため、第三者評価と実運用の報告が増えれば評価が変わる可能性がある。

参考文献

- [1] Anthropic 「Introducing Claude Haiku 5.5」 （2026 年 10 月 7 日） . <https://www.anthropic.com/claude-haiku-5-5>

- [2] Anthropic, Claude Platform Docs 「Claude Platform release notes」 (2026 年 10 月 7 日の項) .
<https://platform.claude.com/docs/en/release-notes/overview>
- [3] Anthropic, Claude Platform Docs 「Claude Haiku 5.5」 (モデル概要) .
<https://platform.claude.com/docs/en/models/haiku-5-5/overview>
- [4] Anthropic 「Claude Haiku 5.5 System Card」 (2026 年 10 月 7 日) .
<https://www.anthropic.com/document/claude-haiku-5-5-system-card>
- [5] AWS Machine Learning Blog 「Introducing Claude Haiku 5.5 on AWS」 (2026 年 10 月) .
<https://aws.amazon.com/blogs/machine-learning/introducing-claude-haiku-5-5-on-aws/>
- [6] Unite.AI 「Anthropic Releases Claude Haiku 5.5, Cutting Small-Model API Prices」 (2026 年 10 月) .
<https://www.unite.ai/anthropic-releases-claude-haiku-5-5-cutting-small-model-api-prices/>
- [7] Artificial Analysis 「Claude Haiku 5.5 (max with fallback) - Intelligence, Performance & Price Analysis」 .
<https://artificialanalysis.ai/models/claude-haiku-5-5>
- [8] Artificial Analysis 「Claude Haiku 5.5 (high with fallback) - Intelligence, Performance & Price Analysis」 .
<https://artificialanalysis.ai/models/claude-haiku-5-5-high>
- [9] Artificial Analysis 「Claude Haiku 5.5 (medium with fallback) - Intelligence, Performance & Price Analysis」 .
<https://artificialanalysis.ai/models/claude-haiku-5-5-medium> ※ページ本体は取得できず (404) 。値は
検索結果の抜粋による (要確認)
- [10] Codersera 「GPT-6 Sol & Luna: Pricing, Benchmarks, vs Astra」 . <https://codersera.com/blog/gpt-6-sol-luna-complete-guide-2026/>
- [11] Shattered.io 「Claude 5.5 Family: 2 Models in 6 Days, Haiku Next [2026]」 . <https://shattered.io/claude-5-5-family-2-models-6-days-haiku-next-2026/>
- [12] CellCog 「Claude Haiku 5.5 Release Date: The Haiku 5 That Never Came」 . <https://cellcog.ai/blog/claude-haiku-5-5-release-date/>
- [13] Emergent 「Claude Haiku 5.5 Release Date: What Anthropic Has Confirmed」 .
<https://emergent.sh/news/claude-haiku-5-5-release-date>
- [14] QCode 「Claude Haiku 5.5: release timing and what is known」 . <https://qcode.cc/en/claude-haiku-5-5-status-tracker>
- [15] OrcaRouter 「Claude Haiku 5.5 Is Coming, Weeks After a Leak Said It Died」 .
<https://www.orcarouter.ai/blog/claude-haiku-5-5-leak>
- [16] Anthropic, Claude Platform Docs 「Claude Haiku 4.5」 .
<https://platform.claude.com/docs/en/models/haiku-4-5/overview>
- [17] Anthropic, Claude Platform Docs 「What's new in Claude Haiku 5.5」 .
<https://platform.claude.com/docs/en/models/haiku-5-5/whats-new-haiku-5-5>
- [18] Jake Handy 「Model Drop: Claude Haiku 5.5」 Handy AI (Substack, 2026 年 10 月) .
<https://handyai.substack.com/p/model-drop-claude-haiku-55>
- [19] GitHub Issue #220 「Add Claude Haiku 5.5 translation via Anthropic and OpenRouter」

- (kapitalismho/PuriPuly-heart) . <https://github.com/kapitalismho/PuriPuly-heart/issues/220>
- [20] Kingy AI 「Claude Haiku 5.5: Benchmarks, Specs, Pricing and How It Compares with GPT-6 Sol and Luna」 . <https://kingy.ai/blog/claude-haiku-5-5-benchmarks-specs-sol-luna/>
- [21] Artificial Analysis 「GPT-6 Luna (max) - Intelligence, Performance & Price Analysis」 . <https://artificialanalysis.ai/models/gpt-6-luna>
- [22] Kingy AI 「GPT Sol & Luna vs Claude Haiku 5.5: Specs and Benchmarks」 . <https://kingy.ai/blog/gpt-sol-luna-vs-claude-haiku-5-5/>
- [23] OrcaRouter 「Gemini 3.8 Flash vs Claude Haiku 4.5: which fast model」 . <https://www.orcarouter.ai/blog/gemini-3-8-flash-vs-claude-haiku-4-5>
- [24] Rama Digital 「Luna vs DeepSeek Flash vs Gemini Flash vs Haiku: Price & Use Case」 . <https://ramadigital.id/en/blog/gpt-6-luna-vs-deepseek-flash-vs-gemini-flash-vs-claude-haiku>
- [25] Artificial Analysis 「Claude Sonnet 5.5 vs Claude 4.5 Haiku - Release Intelligence, Performance & Price Comparison」 . <https://artificialanalysis.ai/models/releases/comparisons/claude-sonnet-5-5-vs-claude-4-5-haiku>
- [26] Hacker News 「Claude Haiku 5.5」 スレッド* (2026 年 10 月 7 日) . <https://news.ycombinator.com/item?id=49996437> ※スレッド本体は取得エラー。内容は検索抜粋および二次引用による (要確認)
- [27] Artificial Analysis 「GPT-6.1 Sol (Medium) vs GPT-6 Luna (Max): Model Comparison」 . <https://artificialanalysis.ai/models/comparisons/gpt-6-1-sol-medium-vs-gpt-6-luna>
- [28] Artificial Analysis 「Google has released Gemini 3.8 Flash, its fourth Flash model in under four months」 . <https://artificialanalysis.ai/articles/gemini-3-8-flash>
- [29] Anthropic 「Claude Haiku」 (製品ページ) . <https://www.anthropic.com/claude/haiku>
- [30] Anthropic, Claude Code Docs 「Create custom subagents」 . <https://code.claude.com/docs/en/sub-agents>
- [31] Joe Cotellesse 「Claude Code Explore Agent Doesn't Run Haiku. Mine Ran Opus.」 . <https://joecotellesse.com/posts/claude-code-explore-agent-haiku/>
- [32] mirin.pro 「36% of Your Claude Code Subagents Are Running on Haiku」 . <https://mirin.pro/blog/claude-code-subagents-haiku-telemetry/>
- [33] Neural LLM 「Claude Code Subagents on Haiku: What Telemetry Shows — and How Power Claude Makes It Visible」 . <https://neural-llm.com/blog/guides/claude-code-subagent-haiku-telemetry-power-claude>
- [34] Simon Willison 「Introducing Claude Haiku 5.5」 (2026 年 10 月 7 日) . <https://simonwillison.net/2026/Oct/7/claude-haiku-5-5/>
- [35] 窓の杜 「Anthropic、『Claude Haiku 5.5』を発表 ～1 年ぶりの小型モデル、最速・最安を更新」 . <https://forest.watch.impress.co.jp/docs/news/2146582.html>
- [36] ビジネス+IT 「アンソロピック、小型 AI 『Claude Haiku 5.5』 発表 入力 100 万トークン 0.10 ドルから」 . <https://sbbbit.jp/article/cont1/187342>

- [37] daily.dev 「Anthropic releases Claude Haiku 5.5, its newest small model」 .
<https://daily.dev/posts/anthropic-releases-claude-haiku-5-5-its-newest-small-model-wli6uk40j>
- [38] DEV Community 「Haiku 5.5 Cut Prices 90%, Then the Fine Print Showed Up」 .
<https://dev.to/dishant0406/haiku-55-cut-prices-90-then-the-fine-print-showed-up-5apn>
- [39] Yahoo Finance 「Anthropic reveals Haiku 5.5 model as AI pricing war intensifies」 .
<https://finance.yahoo.com/technology/article/anthropic-reveals-haiku-55-model-as-ai-pricing-war-intensifies-180000423.html>
- [40] Benzinga 「Anthropic Takes Aim At AI Costs With Claude Haiku 5.5」 .
<https://www.benzinga.com/markets/private-markets/26/10/62231753/anthropic-takes-aim-at-ai-costs-with-claude-haiku-5-5>
- [41] Cemri, M. ほか 「Why Do Multi-Agent LLM Systems Fail?」 arXiv:2503.13657.
<https://arxiv.org/abs/2503.13657> ※本文書作成時に再取得していない（URL は arXiv 識別子から構成、未検証）
- [42] Anthropic Engineering 「How we built our multi-agent research system」 （2025 年 6 月 13 日） .
<https://www.anthropic.com/engineering/multi-agent-research-system> ※本調査で再取得していない（URL 未検証）
- [43] 日本アイアール 「生成 AI で特許調査はどこまでできる？化学系調査・スクリーニング編」 .
https://nihon-ir.jp/generative-ai_patent-search_chemical1/
- [44] AI JapanStep 「特許調査の精度で汎用 LLM を圧倒。知財特化型 AI の実力」 （2026 年 3 月） .
<https://ai.japanstep.jp/learn/2026/03/1728/>